Census of Marine Life
DNA Barcoding Protocol

This is an evolving protocol that will be updated regularly. Comments are welcome.

Purpose. The purpose of this protocol is to encourage investigators working with the Census of Marine Life to determine **DNA barcodes** of collected specimens. DNA barcoding of CoML specimens will provide immediately useful results and will have long-term applications for CoML investigators and the general scientific community.

Background and Rationale. DNA sequence analysis of a uniform target gene to enable species identification has been referred to as **DNA barcoding**, by analogy with the UPC bar codes used to identify manufactured goods (1). A remarkably short DNA sequence should contain more than enough information to distinguish 10 or even 100 million species. For example, a 600-nucleotide segment of a protein-coding gene contains 200 codon third nucleotide positions. At these sites, substitutions are (usually) selectively neutral and mutations accumulate through random drift. Even assuming that a group of organisms is completely biased to AT or GC at third nucleotide positions, there are then 2 possible bases at 200 positions, or $2^{200} = 10^{60}$ possible sequences based on third nucleotide positions alone. Proof of principle for DNA barcoding has been provided by comparison of mitochondrial cytochrome c oxidase subunit I (COI) sequences among closely related species and across diverse phyla in the animal kingdom (2).

DNA barcoding has the potential to be a practical method for identification of the estimated 10 million species of eukaryotic life on earth. As a uniform method for species identification, DNA barcoding will have broad scientific applications. It will be of great utility in conservation biology, including biodiversity surveys. It could also be applied where traditional methods are unrevealing, for instance identification of eggs and immature forms, and analysis of stomach contents or excreta to determine food webs. In addition to enabling species identification, DNA barcoding will aid phylogenetic analysis and help reveal the evolutionary history of life on earth.

An appropriate target gene for DNA barcoding is conserved enough to be amplified with broad-range primers and divergent enough to allow species discrimination. The initial target for DNA barcoding described in this protocol is **mitochondrial cytochrome c oxidase subunit I (COI)**. Selection of an appropriate gene is a critical strategic and practical decision, with significant consequences for the overall success of this project. A number of genes may be likely to meet one or more of our goals (discrimination and identification of species, discovery of new and cryptic species, reconstruction of evolutionary relationships among species and higher taxa). Our selection of COI as a target gene is supported by published and ongoing work, which demonstrates that barcoding via COI will meet the project goals for a wide diversity of animal taxa (1-3).

Limitations. An important outcome of this project will be to identify the groups in which alternate targets are needed and to define what those targets should be. Cnidarians (sea anemones, corals, and some jellyfish) for example, have little mitochondrial sequence diversity, perhaps due to a supplemental mitochondrial DNA repair system, and a nuclear gene target will likely be needed. Recently diverged species and species that have arisen through hybridization may not be resolved by COI sequencing. Information regarding the success or lack thereof of COI barcoding is welcome and will be incorporated into this site.

This protocol addresses DNA barcoding of animal species. Alternate targets or protocols will likely be needed for DNA barcoding of other eukaryotes. Plants have too little mitochondrial sequence diversity, probably due to hybridization and introgression (potential targets include matK, a chloroplast gene, and ITS (internal transcribed spacer), a nuclear gene). The mitochondrial DNA of fungi contains introns, which can complicate DNA amplification (this could be circumvented by applying RT-PCR). For protists, the utility of COI sequencing has not been explored in depth.

Methods. The essential points are 1) specimen preservation in 95% ethanol (not formalin) to facilitate DNA isolation, 2) amplification and sequencing of uniform target gene(s), and 3) databasing of DNA sequences linked to specimens including ancillary data. We welcome information regarding useful techniques (e.g. ZooGene protocol for zooplankton - http://www.zoogene.org/main/sample_preservation_protocol.html) (3).

1. Specimen Preservation. To allow DNA isolation, 95% ethanol should be used—DNA is difficult to extract from formalin-preserved specimens. The ethanol should generally be poured off and replaced with new 95% ethanol within a few days of collection to optimize DNA preservation.

DNA has been successfully extracted from formalin-preserved tissue, including relatively ancient samples and these techniques may be important in examining previously archived specimens (4,5).

2. Specimen labeling. The usefulness of DNA barcoding depends on linking the sequence to a specimen and its associated data (collector, taxonomic confirmation, date, georeference coordinates, etc.).

3. DNA isolation. In addition to standard methods, there are commercial kits (e.g. Sigma-Aldrich product number GDI-3) that are inexpensive and have high success in recovering DNA.

4. Gene target(s). The initial target sequence is the 5' segment of COI, as described below.

Barcoding Basic: Mitochondrial cytochrome c oxidase subunit I, 5' segment (COI-5'). Broad range primers are available that will amplify an approximately 700-bp segment from diverse invertebrates (including Annelida, Arthropoda, Coelenterata, Echinodermata, Echiura, Mollusca, Nemertina, Platyhelminthes, Pogonophora, Sipuncula, and Tardigrada) (6):

5’ primer (5’-ggtcaacaaatcataaagatattgg-3’)
3’ primer (5’-taaacttcagggtgaccaaaaaatca-3’)

These primers will also amplify COI-5’ from some chordates. Additional primers, with broad application for chordates and GC-rich species, are under evaluation and will be added to this site.

If COI-5’ is not sufficient for species discrimination, other rapidly evolving gene(s) may need to be analyzed as potential barcoding targets. Possible supplementary sequences include the complete COI gene, other mitochondrial genes (e.g. 16S rRNA, cytochrome b), and/or ITS (internal transcribed spacer), which is a nuclear gene located between rRNA genes.

Barcoding Elective I: Small subunit nuclear ribosomal RNA (SSU rRNA). SSU rRNA, also referred to as 18s rRNA, is a slowly evolving gene useful for deeper phylogenetic analysis. In addition to the analysis of COI-5’, we encourage CoML investigators to determine SSU rRNA sequences from specimens. SSU rRNA is the basis for the Tree of Life and other comprehensive examinations of evolution of life. The following primers will amplify SSU rRNA (approximately 1800 bp) from most eukaryotes (7):

5’ primer (5’-aacctggttgatcctgccagt-3’)
3’ primer- (5’tgatccttctgcaggttcacctac-3’)

5. Databasing results. CoML Barcode sequences should be submitted to GenBank (http://www.ncbi.nlm.nih.gov/GenBank). We encourage investigators to post results to the Ocean Biogeographic Information System (OBIS – http://www.iobis.org), including a link on OBIS species pages to the relevant sequence deposited in GenBank. In addition, a Barcode of Life database will soon be launched under the auspices of the Hebert laboratory that integrates sequence data with taxonomic and specimen information.

References

1. Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological Identifications through DNA barcodes. Proc R Soc Lond B 270:313-322.
2. Hebert PDN, Ratnasingham, deWaard JR. 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc Lond B (in press).
3. Bucklin A, Frost, BW, Braqdfor-Grieve J, Allen LD, Copley NJ. 2003. Molecular systematic and phylogenetic assessment of 34 calanoid copepod species of Calanidae and Clausocalanidae. Mar Biol 142:333-343.
4. Fang SG, Wan QH, Fijihara N. 2002. Formalin removal from archival tissue by critical point drying. BioTechniques 33:604-611.
5. France SC, Kocher TD. 1996. DNA sequencing of formalin-fixed crustaceans from archival research collections. Mol Mar Biol Biotechnol 5:304-313.

6. Folmer O, Black M, Hoeh W, Lutz R, Vrigenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3:294-299.
7. Medlin L, Elwood HJ, Stickel S, Sogin ML. 1988. The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene 71:491-499.

Resources

The Paul Hebert Laboratory website (http://www.uoguelph.ca/~phebert/) includes an informative explication of barcoding and links to pdf files of published works.

The European Ribosomal RNA Database (http://oberon.rug.ac.be:8080/rRNA/) is a curated database that contains, in addition to sequence data, information about primers, phylogenetic trees, and software for sequence alignment and tree construction.

The ZooGene site (http://www.zoogene.org/) contains information on zooplankton (calanoid copepods and euphausiids) genomics, including utility of COI sequencing and protocols for specimen preparation and DNA isolation.

Acknowledgements. The genesis of this protocol is a “Taxonomy and DNA” conference held March 9-12, 2003 at Cold Spring Harbor Laboratory, New York. The conference participants included CoML investigators and other specialists in animal, microbial, and plant taxonomy, molecular biology, and informatics. This protocol was prepared by Ann Bucklin, Paul Hebert, Nancy Knowlton, and Mark Stoeckle.

Comments, additions, and/or revisions are welcome (**contact** MarkStoeckle@nyc.rr.com).